Transferring Files Between AWS S3 and Job-Time Filesystems

Use these **CLOUD** directives to copy files between your job-time filesystems on AWS and your S3 storage on AWS.

```
#CLOUD -volume_put=s3_folder_name_to_copy_to
#CLOUD -volume_get=s3_folder_name_to_copy_from
```

Please note the following:

- All **CLOUD** directives should be placed after the PBS directives in a PBS script.
- The **-volume_put** and **-volume_get** directives should be placed after the **â volume_type** directive in a PBS script, and can be placed either before or after the **â volume_size** and **â volume_mount** directives of the corresponding job-time filesystem.
- No leading slash should be used in the **s3_folder_name**, as demonstrated in the example in this section.
- The S3 folders are all relative to your S3 root location (i.e., /). Other users, even those in the same group, cannot see your files in S3.
- The **-volume_get** directives will get everything inside the specified folder. Getting a single file from an S3 folder is not currently supported.

If you give the **-volume_put** directive for a job-time filesystem where multiple nodes have their own space (such as the *ephemeral* or *local* type of filesystem, but not the *shared* type), then files from all of the nodes will be put in the same S3 folder. This results in files with the same name on each node ending up in an undetermined state as to which node's version is stored. You can avoid this by including **{node}** in the S3 folder name, so the path will have **nodeX** appended to it when saved on the *X* node's copy. For example, for a job that asks for three nodes with these directives:

```
#CLOUD -volume_type=local
#CLOUD -volume_put=run12/{node}
#CLOUD -volume_mount=/data
```

the following directories in your S3 environment will be created and files in the **/data/** directory from each node will be stored in the appropriate one:

```
/run12/node0
/run12/node1
/run12/node2
```

The **-volume_get** directive works similarly in these situations, but without the need for the **{node}** text.

For a filesystem type that is unique to each node (for example, ephemeral, local, headnode, or node=*X*), if the number of nodes used in the 'put' operation is less than the number of nodes used in the 'get' operation, then some nodes may have empty directories from the 'get' operation. However, if the number of nodes used in the 'put' operation is more than those in the 'get' operation, then not all the data from the 'run12' S3 bucket will be brought back.

For more information, see:

- Examples of Job-Time Filesystem Related Directives
- Optional Job-Time Filesystem Related Directives

Order of Operation

- The **-volume_get** operation happens when the compute instance boots.
- The **-volume_put** operation for shared filesystems occurs after the job exits, and therefore will not show up in the output of **â volume_list** (if included in the job script). You can use the **nas_s3_ls** command (described in this article) to see them in S3 afterwards.
- The **-volume_put** operation for non-shared filesystems occurs after the PBS script but before the **â volume_list** operation.

If you are using other **CLOUD** directives, see Order of Operation of Cloud Directives.

Article ID: 592
Last updated: 25 Jan, 2021
Revision: 24
Cloud Computing -> AWS Cloud -> Transferring Files Between AWS S3 and Job-Time Filesystems
https://www.nas.nasa.gov/hecc/support/kb/entry/592/